The Dissertation Committee for Nahar Albudoor Certifies that this is the approved version of the following Dissertation:

Identifying Language Disorder in Bilingual Children Using Automatic Speech Recognition: A Feasibility Study

Committee:

Rajinder Koul, Supervisor

Elizabeth Peña, Co-Supervisor

Lisa Bedore

Chang Liu

Craig Champlin

Identifying Language Disorder in Bilingual Children Using Automatic Speech Recognition: A Feasibility Study

by

Nahar Albudoor

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2021

Acknowledgements

I would like to express my sincerest gratitude to the many people who made this endeavor possible:

To Elizabeth Peña – thank you for your invaluable mentorship, kindness, and patience. Your wealth of knowledge, incredible drive, and commitment to this field continually drove me to be a better researcher.

To Lisa Bedore, Rajinder Koul, Craig Champlin, and Chang Liu – many thanks for guiding me on this path with your expertise, advice, and support.

To my friends and colleagues at the HABLA Lab – thank you for the conversations, the laughs, and the tears, but most importantly, thank you for being generous with your knowledge and passion.

To the children and families who participated in this study – my heartfelt thanks for your time and effort. Without your willingness to complete these often arduous tasks, this research would not be possible.

To my sisters, who are not only my sisters but my best friends. Thank you for being there, whether it was when you were my roommates or when you were an ocean away. Your unconditional love, understanding, and support during this time meant the world to me.

To my father – thank you for being my rock. You never wavered in your encouragement of my endeavors. You always stood by side, lending a helping hand and a listening ear. For that I am immensely grateful.

To my mother – you have been my hero for as long as I can remember and every day that passes, that sentiment only grows. I hope to inspire the amount of pride in you that you have inspired in me. Thank you for showing me fortitude, dedication, and love.

To Jissel – there are no bounds to my gratitude to you. You have been my partner through this in every sense of the word. Your dogged belief in me along with your own remarkable insight and passion have inspired me not only to be a better academic, but also to be the best version of myself. I am so grateful to have you in my life.

Abstract

Identifying Language Disorder in Bilingual Children Using Automatic Speech Recognition: A Feasibility Study

Nahar Albudoor, Ph.D.

The University of Texas at Austin, 2021

Supervisors: Rajinder Koul and Elizabeth Peña

The differential diagnosis of developmental language disorder (DLD) in bilingual children represents a unique challenge due to their distributed language exposure and knowledge. The current evidence indicates that dual-language testing yields the most accurate classification of DLD among bilinguals, but there are limited personnel and resources to support this practice. This study explored the feasibility of dual-language automatic speech recognition (ASR) for identifying DLD in bilingual children. Eighty-four Spanish-English bilingual second graders with (n = 25) and without (n = 59) confirmed diagnoses of DLD completed the Bilingual English-Spanish Assessment – Middle Extension (BESA-ME) Morphosyntax in both languages. Their responses on a subset of items were scored manually by human examiners and programmatically by a researcher-developed ASR application employing a commercial speech-to-text algorithm. Results demonstrated moderate overall item-by-item scoring agreement (k = 0.54) and

similar diagnostic accuracies (human = 92%, ASR = 88%) between the two methods using the best-language score. Classification accuracy of the ASR method increased to 94% of cases correctly classified when test items with poorer discrimination in the ASR condition were eliminated. These findings establish the concurrent validity of the BESA-ME Morphosyntax for Spanish-English bilingual second graders when ASR is used to process their responses. More broadly, this study provides preliminary support for the technical feasibility of ASR as a bilingual expressive language assessment tool.

Table of Contents

List of Tables

Chapter 1: Background

There is a critical need for valid language measures for the 12 million bilingual children in the United States (U.S. Census Bureau, 2019). Bilingual and other linguistically diverse children are disproportionately over- and underidentified with developmental language disorder (DLD) (Morgan et al., 2015; Sullivan & Bal, 2013), an impairment in the comprehension and production of language that affects one in 10 children (Norbury et al., 2016). The consequences of DLD misdiagnosis have significant societal impact. Underidentification of DLD is associated with increased academic failure, dropout, and incarceration (Anderson et al., 2016; Thurlow & Johnson, 2011), risks that disproportionately affect persons who are minoritized (Pettit & Western, 2004; Wood et al., 2017). Overidentification of DLD compromises resource allocation and contributes to the increased marginalization of language minority students who receive diagnoses of DLD simply because they are bilingual.

Best practices for assessing bilingual children require the evaluators to consider both languages of a child (American Speech-Language-Hearing Association, 2021; Royal College of Speech and Language Therapists, 2021). However, only 6% of U.S. speech-language pathologists (SLPs), the practitioners qualified to diagnose DLD, are bilingual (American Speech-Language-Hearing Association, 2018) relative to the 22% of bilinguals in the U.S. population (U.S. Census Bureau, 2019). To address this disproportionality, it is critical to equip all SLPs with tools for assessing bilingual children, even when the SLPs are monolingual or bilingual in a language pair that differs from the child's. Automatic speech recognition (ASR), the technology that processes human speech and converts it to text, holds promise for meeting this need. Multipurpose consumer ASR systems, such as Google Assistant and Amazon Alexa,

can process up to 120 human languages and variants and their algorithms are available for use in custom applications, while custom-developed ASR systems can be trained and programmed for specific use cases (Sabu & Rao, 2018). An ASR system that achieves a sufficient level of transcription agreement with proficient speakers of a target language may extend access to that language to all SLPs, enabling them to obtain information about children's skills in languages they do not speak. While this technology would not take the place of a comprehensive evaluation by a practitioner who speaks both or all a child's languages, it may serve as an indicator of that child's risk for DLD.

There is evidence, both longstanding and emerging, that ASR is technically and practically feasible for assessing children's language skills. Dictation software, which converts speech to text, has long been used as an academic accommodation for children with special education needs (Bolt & Thurlow, 2004; Thompson et al., 2002) and has been shown to achieve a mean word-level accuracy of 87% (MacArthur & Cavalier, 2004). Newer educational software using ASR and artificial intelligence conducts more advanced conversational exchanges with children as successfully as human interlocuters, including delivering prompts such as a story comprehension questions, processing children's oral responses to the prompts, and responding appropriately (Xu et al., 2021). Still, ASR has yet to be employed in the assessment of children's language skills and the extent to which ASR can be used to identify DLD is currently unknown. Thus, the purpose of the present study was to provide an early indication of the extent to which ASR may be used for DLD identification. A researcher-developed ASR application was employed to transcribe the responses of Spanish-English bilingual second graders who completed a validated morphosyntax assessment task in Spanish and English. Children were in two groups, those with DLD and those with typical language development (TD). Children's

ASR-transcribed responses were programmatically scored and analyzed for agreement with human scorers and for DLD classification accuracy in order to determine the feasibility of ASR for identifying DLD in bilingual children.

Identifying Developmental Language Disorder in Bilingual Children

Children who speak a language other than English in the home represent nearly one quarter of all children in the U.S. (U.S. Census Bureau, 2019), but there remain relatively few measures that are validated for the identification of language disorders in this population. The principal reason for this is that the detection of language difficulties among bilingual children represents a unique challenge. Bilingual children's experiences are heterogenous, with their linguistic environments varying as a function of factors such as their geographic location, age of first exposure to another language, first language prestige, family structure, family language history, and academic program, among others (for review, see De Houwer, 2018). This variation results in fluctuations in language exposure that are ultimately associated with variations in language performance. There are two primary features of these variations that can affect the detection of DLD. First, while bilingual children's language skills generally reflect their exposure to each language (Allen et al., 2002; Blom, 2010; Hoff et al., 2012; Meisel, 2007), it is typical for bilingual children to exhibit mixed or shifting patterns of language performance and dominance (Bedore et al., 2012; Bedore et al., 2010; Lugo-Neris et al., 2015; Oppenheim et al., 2020; Rojas & Iglesias, 2013). For example, in a longitudinal study of typically developing Spanish-English bilingual children by Rojas and Iglesias (2013), children's English narrative skills between kindergarten and the end of second grade followed a curvilinear, rather than linear, trajectory. While the overall change was positive, showing that there was general growth in children's English language skills, there were marked periods of decreased performance that

occurred during children's summer breaks, when they were exposed to less English. These findings suggest that measuring bilingual children's skills in a single language at a single time point may lead to the erroneous assumption that their development is atypical. Second, bilingual children with typical language development can make errors in one or both of their languages that are like errors made by monolingual children with DLD (e.g., Gutiérrez-Clellen et al., 2008; Orgassa & Weerman, 2008; Paradis & Crago, 2000). For example, Gutiérrez-Clellen and colleagues (2008) demonstrated that typically developing English language learners had similar rates of errors as English monolinguals with DLD in their production of English past tense -ed, third person singular -s, and present tense auxiliary BE. Such findings indicate that children who are in the process of acquiring a second language can present language characteristics that are like monolingual children with DLD, making differential diagnosis more challenging.

The practical effect of these features on the identification of DLD is that considering only one of a bilingual child's languages generally does not yield a linguistic profile complete enough to determine whether that child's language development is atypical. While alternatives to dual-language testing exist, including methods that involve caregiver report (Pratt et al., 2020) and single-language item sets that are particularly sensitive to DLD errors (Gutiérrez-Clellen et al., 2006), the most validated practice for detecting DLD in bilingual children is to consider both languages (for review, see Ebert & Kohnert, 2016). This is because when performance in both languages is considered, bilingual children with DLD consistently demonstrate deficits across languages that confirm a differential diagnosis of DLD, while bilingual children with typical language development demonstrate typical performance in their better language. For example, Thordatottir and colleagues (2006) showed that, when tested in each language separately, typically developing French-English bilingual preschoolers' expressive vocabulary, receptive

syntax, and expressive syntactic skills were significantly lower in either language than those of French or English monolingual preschoolers. When performance across languages is considered, however, bilingual children with typical language development are distinct from those with DLD. A series of studies by Peña and colleagues (Anaya et al., 2016; Peña & Bedore, 2011; Peña et al., 2016; Peña et al., 2020) demonstrated that when Spanish-English bilingual children between preschool and fourth grade were tested in both of their languages, their coordinate scores (i.e., the score in the test language in which they performed better) consistently produced the most accurate classifications of DLD and TD. These findings illustrate the distinction between language differences due to typical bilingual experiences and those due to language disorders, highlighting the need for testing in both languages. Such research has led to policy statements that emphasize the assessment of both languages of a bilingual child, such as those issued by the American Speech-Language-Hearing association (2021) and the Royal College of Speech and Language Therapists (2021).

Barriers to Access

Despite the consensus on the importance of dual-language testing of bilingual children, there are significant barriers to its implementation that can contribute to the over- or underidentification of DLD in this population. To collect information about two languages, evaluating SLPs must either be proficient in both languages or they must establish a method of obtaining information about the language they do not speak. Among the small percentage of SLPs who are bilingual, challenges include lack of time and access to assessment materials/training are cited as common barriers to bilingual testing (Arias & Friberg, 2017). Among SLPs who are not bilingual, methods such as collaborating with interpreters also have challenges. For example, Saenz and Langdon (2019) found that of the 208 California SLPs they

surveyed who reported previously working with interpreters, 40% expressed uncertainty about the training of the interpreters. Specifically, 60% reported working with adult family members or friends and 27% reported working with family members who were minors for interpreting. Furthermore, most SLPs (60%) reported that they had experienced instances when they needed to work with an interpreter but could not, most commonly due to an inability to find one (69%), uncertainty about the interpreter's training (26%), lack of necessary assistance from the interpreter (23%), or lack of monetary support from an employer (16%). Such challenges represent barriers to the timely and accurate diagnosis of DLD among bilinguals that increases their risk for DLD misdiagnosis. This can later intersect with other risk factors affecting this population. In the U.S., both under- and overidentification of speech and language impairments have been documented among linguistically diverse children. In a study of 17,837 children in the U.S. with a mean age of 10.2 years (SD = 3.6 years), among whom 16.8% were identified as limited English proficient (LEP), Sullivan and Bal (2013) observed that LEP children were 28% more likely to be identified with speech or language impairments than their English-speaking peers. When Morgan and colleagues (2015) attempted to replicate these findings in a longitudinal analysis of 20,100 U.S. elementary and middle schoolers, they observed a reverse pattern of misdiagnosis when they adjusted for covariates. LEP children in their study were 40% less likely to be identified with speech or language impairments than otherwise similar English-speaking children. The contradictory findings of these studies illustrate the nuances associated with language disorder identification among bilingual and other linguistically diverse children.

Critically, phenomena that disproportionately affect minoritized populations, such as academic dropout and incarceration (Pettit & Western, 2004; Wood et al., 2017), also disproportionately affect children with language disorders (Anderson et al., 2016; Thurlow &

Johnson, 2011). A systematic review of 17 studies examining juvenile offenders in the U.S., United Kingdom, and Australia confirmed that the rate of language disorder was significantly higher among juvenile offenders than among matched non-offenders across all studies, without exception (Anderson et al., 2016). Furthermore, the rate of high school dropout among U.S. students with speech or language impairments is 8.4%, more than double that of the national average of 4.1% (Thurlow & Johnson, 2011). These findings indicate that bilingual and other linguistically diverse children experience overlapping risk factors that can amplify long-term challenges, further underscoring the need for accurate DLD identification practices for this group.

Automatic Assessment to Expand Access to Accurate Identification

Recent work has explored the use of automatic language tasks as a potential alternative to person-administered dual-language assessment. For example, de Villiers and colleagues (2021) reported on the development of the Quick Interactive Language Screener: English-Spanish (QUILS: ES), an electronic language screening instrument that automatically administers and scores receptive vocabulary, syntax, and processing tasks in both English and Spanish using a touchscreen tablet interface. The QUILS:ES results of Spanish-English bilingual children aged 3 to 5;11 were significantly correlated with their results on the English (r =.693, $p < .001$) and Spanish (r = .449, $p < .002$) Preschool Language Scales, Fifth Edition (PLS-5), providing evidence for the instrument's concurrent validity. Furthermore, the instrument produced good internal consistency (English = 0.89, Spanish = 0.85) and test-retest reliability (0.89), indicating good fit of the test items with the measurement constructs and moderately high correlations between measurement occasions, respectively. Similarly, pilot analyses from Pratt and colleagues (in review) explored the feasibility of a remote, computer-administered English and

Spanish language assessment protocol that employed prerecorded test items from the Bilingual English Spanish Assessment (Peña et al., 2018) to measure children's oral language skills. In response to receptive test items, children of certified SLP parents listened to the prompts and clicked on their responses from picture arrays, which were automatically scored. In response to expressive test items, children listened to the prompts, answered verbally, and their responses were manually scored by the evaluators. Evaluators were the children's SLP parents, who administered half of the test items remotely (from a different room of the house) and half of the test items in person. Results from the split-half administrations demonstrated a significant correlation ($r_s = .979$, $p < .01$) between children's results in the remote condition and their results in the in-person condition, demonstrating the instrument's concurrent validity.

These previous studies represent an emerging area of research in bilingual DLD assessment that holds considerable promise for extending access to multiple languages to all SLPs. That is, de Villiers and colleagues (2021) and Pratt and colleagues (in review) provided initial proof of concept for the use of automatically narrated and scored test items to assess children's oral language skills in more than one language. A primary limitation of both works, however, was that automatic scoring was limited to receptive test items. While the evaluation of receptive language may be sufficient for screening purposes (in the case of the QUILS: ES, for example), instruments used for DLD diagnosis consistently elicit children's expressive language skills. For example, among bilingual DLD identification measures, all measures with fair to good classification accuracy (i.e., sensitivity and specificity above 80%) reviewed by Brinson and colleagues (2020) contained expressive production tasks or test items. Although the purpose of the Brinson and colleagues (2020) review was not to compare receptive and expressive tasks, their findings demonstrated that expressive tasks represent a core aspect of bilingual DLD

assessment at present. As such, the current evidence provides preliminary technical and practical support for the automatic scoring of receptive language tasks, but further research is necessary for establishing the automatic scoring of expressive language tasks. Such work is a critical step toward the development of instruments that are fully automatically administered and scored and is therefore the focus of the current study.

AUTOMATIC RECOGNITION OF CHILDREN'S SPEECH

Technical Considerations

Considerable evidence indicates that ASR is technically viable for processing children's speech for the purposes of expressive language assessment. Dictation through ASR software has existed as an academic accommodation for children receiving special education services from as early as 1997, when Dragon Systems released the first computer software that converted connected speech from audio to text (Bolt & Thurlow, 2005; Dragon Systems, Inc, 1997; Thompson et al., 2002). An early study of 14-year-olds with and without learning disorders (LD) who completed a sentence probe task using the Dragon: Naturally Speaking software demonstrated that the software produced an overall word accuracy rate of 87%, with no significant difference in accuracy between the LD and non-LD groups (MacArthur & Cavalier, 2004). These results provided initial evidence that ASR systems could capture the speech of children, even those with special needs. More recently and with younger children, including those with speech production deficits, researcher-developed ASR systems have achieved even higher accuracy rates. In a study of seven- to nine-year-old children with diagnosed speech sound disorders, Hair and colleagues (2019) tested multiple trained ASR models for their accuracy for analyzing the children's speech at the single-word level. The best-performing ASR model achieved a mean 90% accuracy analyzing children's speech at the single-word level. At

the sentence level, Sabu and Rao (2018) developed an ASR system that achieved a word error rate (i.e., the percentage of recognized words containing substitutions, deletions, or insertions) of just 3.44% when it was used to analyze the speech of 20 students between the ages of 10 and 14 years who completed a sentence oral reading task. These findings indicate that both commercial and custom ASR applications can achieve high accuracy and low error rates when processing children's speech, even when it contains production errors. Although what constitutes "high enough" ASR transcription accuracy is ultimately usage-dependent, the threshold for human word level accuracy for adults with normal hearing is approximately 95% when the signal to noise ratio is above -5 decibels (Spille et al., 2018). As such, some of the child ASR models in the reviewed studies achieved word accuracy rates near or on par with adult human listeners of other adult speakers.

Practical Considerations

The evidence for ASR's accuracy at processing children's speech provides preliminary technical support for the use of ASR in child language applications, including those that may be used for assessment purposes, but such applications must also be practically feasible. That is, a second prerequisite for exploring ASR as a language assessment tool is whether children can be reasonably expected to engage with the technology long enough to complete the necessary tasks. Although work in this area is emerging, there is evidence that even very young children can successfully engage in language elicitation tasks with ASR-embedded devices for up to 30 minutes per session. In a study of three- to six-year-old children, one third of whom were bilingual or English language learners, Xu and colleagues (2021) demonstrated minimal differences in task performance between children who engaged in an English storybook reading and scaffolded comprehension task with an adult (n = 31) and children who engaged in the same

task with a conversational agent embedded in a Google Home Mini device (n = 33). Both the adults and the conversational agent, voiced by an adult woman, told a story that accompanied a physical storybook (common across participants) and presented the children, who were randomly assigned to either condition, with open-ended comprehension questions about the story at spaced intervals throughout the task. Exchanges were conducted manually by the adults, who used a script, but the conversational agent used ASR to process children's responses and dialog programming to provide appropriate feedback. Results showed no significant differences between the two conditions in the accuracy, topical relevance, lexical productivity, and lexical diversity of children's responses unless comprehension questions required high cognitive demand (e.g., inferencing), in which case there was a slight human partner advantage on relevance, productivity, and diversity. Furthermore, there was a slight conversational agent advantage on children's intelligibility. Although the authors did not examine task difficulty, duration, or completion as study variables, they noted that the activity took approximately 20 minutes and that all children in both conditions completed the task, with no children withdrawing from the study due to the demands of the activity. In addition to providing evidence for ASR's processing of multi-word utterances by linguistically diverse children as young as three, these findings demonstrated that it was possible to conduct a lengthy and interactive language task via an automatic agent with this group and that children may even improve their intelligibility with non-human interlocuters. Together with their finding of no negative effect of the agent on the accuracy of children's responses to the comprehension questions, Xu and colleagues' (2021) results provided compelling support for the use of ASR to conduct language tasks and collect information about the correctness of children's responses, a feature critical for analyzing assessment information.

Pilot work by Yeung and colleagues (2019) also showed that children can successfully complete tasks with automatic agents. In their study, 15 preschoolers and 18 kindergarteners interacted with a small, researcher-developed robot. The robot, embedded with a monitor to display images and a speaker to play audio, administered the Sounds in Words and Sounds in Sentences tasks of the Goldman Fristoe Test of Articulation, 3rd Edition (Goldman, 2015), which required children to repeat sentences and expressively identify single words for the purposes of speech error assessment. While the evaluation of ASR to process children's responses to test items was cited as a future goal of the research and therefore not explored in the publication, the authors noted that 80% of the preschoolers and 78% of the kindergarteners completed all test items during the 30-minute experiment. The remaining children discontinued the task due to boredom, fatigue, or in the case of one preschooler, appearing intimidated by the robot. The authors also reported that their greatest challenges were associated with children's difficulty identifying when to begin responding and the audio recordings failing to capture the initial phonemes in children's responses when they contained the article "the" and demonstratives "this" and "that". Yeung and colleagues' (2019) findings highlighted important considerations for the practical implementation of computer- or robot-administered assessment tasks, but they further confirmed that most children, even very young children, could complete such tasks.

Automatic Speech Recognition for Bilingual Language Assessment

The research on both the technical and practical feasibility of ASR for recognizing children's speech indicates that ASR-embedded applications can be used to deliver, process, and score test items for the purposes of language evaluation. However, this research does not demonstrate the extent to which this applies to languages other than English and for children with DLD. A pilot analysis of Spanish-English bilingual second graders' test audio recordings

demonstrated initial evidence toward this aim (Albudoor et al., 2019). Employing the Google Cloud speech-to-text application programming interface (API), we analyzed the audio recordings of 20 Spanish-English bilingual second graders, 10 with confirmed diagnoses of DLD and 10 with typical language development who were matched by age, sex, maternal education, and percent current English language exposure to peers in the DLD group. The children's audio recordings contained their responses to test items from the morphosyntax and semantics subtests of the Bilingual English Spanish Assessment – Middle Extension (Peña et al., 2010) and the narrative comprehension scale of the Test of Narrative Language (Gillam & Pearson, 2004; Gillam et al., 2010) in both English and Spanish. The ASR's overall item-level agreement with human scoring of these measures was 81% for English items and 84% for Spanish items, indicating that the technology demonstrated moderate agreement with human scorers for assessing both Spanish and English test items. The findings provided evidence for the use of ASR to assess the language skills of Spanish-English bilingual children in that they demonstrated that a reasonable degree of scoring agreement could be achieved with ASR. However, there were three limitations to these pilot analyses. First, scoring agreement was calculated using simple percentage agreement by item, which does not account for the possibility of chance agreements or disagreements like more sophisticated metrics of inter-rater reliability, such as Cohen's kappa (Cohen, 1960). Second, scoring agreement was not comprehensively explored. That is, the degree to which agreement varied by test domain, test item, and child-level factors (such as disorder status and language exposure) was not established, important determinations for establishing ASR's sensitivity to such variations. Third, the classification accuracy of the measure (i.e., the degree to which it correctly classified true DLD and TD cases) was not determined as it was not the aim of the study. Establishing classification accuracy is a necessary step toward determining a measure's feasibility for diagnostic purposes. While high human-ASR scoring agreement may suggest classification accuracy like the original measure's, ASR classification may reveal advantages and disadvantages distinct to the technology and must therefore be independently confirmed. Given these limitations, the current study aims to extend the pilot analyses of Albudoor and colleagues (2019) to further

explore human-ASR scoring agreement and to determine the extent to which an ASR measure can independently and accurately classify children with and without DLD.

CURRENT STUDY

This study aims to determine the scoring agreement and classification accuracy of a Spanish-English expressive morphosyntax task, transcribed using ASR technology and scored programmatically, to provide evidence for its feasibility as an assessment tool. Specifically, the first research aim is to determine the item-level agreement between children's original (i.e., human-scored) scores on a Spanish-English bilingual morphosyntax measure and their scores when ASR transcripts are used to score the same assessment and to explore whether agreement varies due to child-level or item-level characteristics. The second research aim, pertaining to DLD identification, is to determine the degree to which the same morphosyntax measure, analyzed using the ASR transcription and scoring procedure, accurately classifies children with their original TD and DLD classifications. The research questions are:

1) What is the ASR-human agreement on a measure of Spanish-English bilingual children's expressive morphosyntax skills?
    (a) Does agreement vary between test languages, test item types, disorder classifications, or exposure groups?
2) What is the classification accuracy of an ASR-scored measure of Spanish-English bilingual children's expressive morphosyntax skills?
    (a) Is it possible to improve the classification accuracy of the ASR scored measure by item analysis and selection?

Chapter 2: Method

PARTICIPANTS

Participants were 84 Spanish-English bilingual second graders with (n = 25) and without (n = 59) developmental language disorder (DLD) whose data were drawn from a larger longitudinal study (NIH R01-DC010366, PI: Elizabeth Peña). The purpose of the original longitudinal study was to follow bilingual children with DLD and to compare their language exposure and performance to typically developing (TD) matched controls. Children with Spanish-English exposure were recruited from two public school districts in central Texas and were in prekindergarten, first grade, or third grade at study entry. They completed one year of language screening followed by four years of annual language and cognition testing. During the screening year, 1,696 children were screened in English and Spanish using the Bilingual English Spanish Oral Screener (BESOS; Peña et al., 2010a). Children were invited to return for longitudinal testing if they were:

(a) at risk for language disorder (i.e., scored below 85 in their better language on one subtest of the BESOS) or

(b) not at risk for language disorder (i.e., scored 85 or above in their better language on one subtest of the BESOS) but matched peers in the at-risk group on age, sex, socioeconomic status, and language exposure (to serve as controls).

At least two control participants were invited for each at-risk child. Children were not invited if they had a history of focal brain injury, severe social-emotional problems, intellectual disability, autism spectrum disorder, or hearing loss, to control for the potential confounds of these conditions. These criteria resulted in 334 children enrolled in the longitudinal testing.

Of these, 84 children were included in the current analysis because they met the following criteria:

(a) completed the morphosyntax subtest of the Bilingual English-Spanish Assessment, Middle Extension (BESA-ME; Peña et al., 2010d) in both English and Spanish at second grade, and

(b) had complete audio recordings.

Second grade was selected as the analysis year as this was the grade with the greatest number of participants, yielding the largest DLD group for classification analyses. Demographics for participants included in the current study are shown in Table 1. The TD and DLD groups did not significantly differ in age ($t = 0.222$, $p = 0.83$), sex ($\chi^2 = 0.394$, $p = 0.53$), maternal education ($t = -0.790$, $p = 0.43$), first English exposure ($t = 2.034$, $p = 0.05$), or current English exposure ($t = -1.440$, $p = 0.16$).

Table 1. *Participant Demographics*

	TD	DLD	Total
n	59	25	84
Age (in Years) – *M (SD)*	7.9 (0.3)	7.9 (0.4)	7.9 (0.4)
Sex (% female)	47%	40%	45%
Maternal Education[a]	2.5 (1.5)	2.3 (1.5)	2.5 (1.5)
Age of First English Exposure (in Years) – *M (SD)*	2.7 (1.6)	3.4 (1.44)	2.9 (1.6)
Percent Current English Input/Output – *M (SD)*	43.0 (14.4)	38.2 (13.3)	41.5 (14.9)

Note. M = mean, SD = standard deviation, TD = typical language development, DLD = developmental language disorder.
[a]Hollingshead (1975) score, where 1 = less than 7th grade, 2 = junior high (9th grade), 3 = partial high school (10th or 11th), 4 = high school graduate, 5 = partial college (at least one year), 6 = college education, and 7 = graduate degree

Developmental Language Disorder Classification

Identification of DLD was conducted as part of the larger Peña and colleagues study (NIH R01-DC010366, PI: Elizabeth Peña) and used a protocol that required converging evidence across multiple indicators. Specifically, children were classified with DLD if they met four of the five following indicators of impairment:

(a) Parent or teacher concern rating (as measured by the Inventory to Assess Language

Knowledge, Peña et al., 2010e) below 4.2 (out of 5) in both English and Spanish,

(b) BESA-ME Field Test morphosyntax score lower than one standard deviation below the normative mean in both English and Spanish,

(c) BESA-ME Field Test semantics score lower than one standard deviation below the normative mean in both English and Spanish,

(d) BESOS composite score lower than one standard deviation below the normative mean in both English and Spanish, and/or

(e) Test of Narrative Language (TNL; Gillam & Pearson, 2004; Gillam et al., 2010) composite score lower than one standard deviation below the normative mean in both English and Spanish.

MATERIALS

Language Environment Measure

The Bilingual Input Output Survey (BIOS; Peña et al., 2010c) measures parent- and teacher-reported child language exposure. The Home BIOS requires a parent or primary caregiver to report the child's Spanish, English, or bilingual input (i.e., language heard) and output (i.e., language spoken) on an hour-by-hour basis from the child's wake time to the child's sleep time on one typical weekday and one typical weekend day. The School BIOS requires the child's classroom teacher to report the child's Spanish, English, or bilingual input and output on an hour-by-hour basis during a typical academic school day from the child's school arrival time to the child's school departure time. These data are used to calculate an overall percent input/output value for each language. Specifically, the weekly sum is divided by the total sum and multiplied by 100 to yield a percent exposure per source and language. For example, if a child reportedly hears 10 hours of English and 30 hours of Spanish at home per week, that

child's percent home English input is 25% (10 ÷ [10 + 30] * 100) and home Spanish input is 75% (30 ÷ [10 + 30] * 100). This calculation is completed for home input, home output, school input, and school output, which are then averaged to yield an overall percent input/output value for each language.

Reference Measures

The following reference measures were used to identify indicators of impairment for children's original DLD diagnoses. Based on their results on these measures, children were classified with DLD if they met four of the five classification indicators listed above.

Bilingual English Spanish Oral Screener

The Bilingual English Spanish Oral Screener (BESOS; Peña et al., in development) is a language screening for Spanish-English bilingual children between prekindergarten and third grade. Preliminary norming for the BESOS demonstrates sensitivity of 0.80 to 0.93 and specificity of 0.92 to 0.94 (depending on the age group) for identifying language disorder (Pena et al., 2018), which is above the 0.80 cut-off Plante and Vance (1994) designated as "fair" for identifying language disorders in children. The BESOS contains one semantics subtest and one morphosyntax subtest in English and Spanish. The semantics subtests measure children's depth and breadth of word knowledge through structures such as functions, definitions, and analogies. The morphosyntax subtests measure children's morphological and syntactic structures through cloze and sentence repetition items. In English, structures include possessive *'s*, regular/irregular past tense, and passives. In Spanish, structures include object clitics, relative clauses, and subjunctives.

Inventory to Assess Language Knowledge

The Home and School Inventory to Assess Language Knowledge (ITALK; Peña et al., 2010e) measure parent- and teacher-reported child language knowledge, respectively. Parents/caregivers and teachers rate children's vocabulary, speech, sentence production, grammar, and comprehension skills on a scale from 0 to 5 for both Spanish and English. Respondents receive descriptors and examples for each point on the scale in order to select the score that best represents the child's skills. The five scores are then averaged to yield one Home ITALK and one School ITALK score for each language that falls between 0 and 5, with 0 representing no skills and 5 representing extensive skills.

Bilingual English Spanish Assessment – Middle Extension Field Test

The Bilingual English Spanish Assessment – Middle Extension Field Test (BESA-ME; Peña et al., in development-a) Field Test is a dual-language measure intended for use with Spanish-English bilingual children between the ages of 7;0 and 11;6 years (see Bedore et al., 2018). The Spanish and English semantics subtests measure semantics breadth and depth through receptive and expressive item types that evaluate a child's ability to identify word functions, categories, definitions, characteristic properties, analogies, similarities and differences, and associations. The English morphosyntax subtest examines possessive -s, third-person singular, regular past tense, plural nouns, present/past auxiliary + progressive -ing, copula negatives, and passives. The Spanish morphosyntax subtest examines articles, present progressive verbs, direct object clitics and subjunctives. The morphosyntax subtests are divided into cloze and sentence repetition sections. Test items from the cloze task require children to expressively complete sentences with words or phrases containing target morphosyntactic forms. Test items from the sentence repetition task require children to verbally repeat full sentences containing target

morphosyntactic forms. Children are assessed both on their ability to repeat the whole sentence (verbatim scoring) as well as on their ability to repeat individual word and phrase targets from the sentence (target scoring). Preliminary classification analyses for the BESA-ME Field Test demonstrate sensitivity of 1.0 and specificity of 0.87 to 0.95 (depending on the age group).

Test of Narrative Language

The Test of Narrative Language (TNL) English (Gillam & Pearson, 2004) is a published, norm-referenced measure of children's narrative language skills for children between the ages of 5;0 and 11;11. The TNL Spanish (Gillam et al., 2010) is an experimental test identical in structure to the TNL English and for which preliminary norming has been conducted. The tests consist of two scales, Narrative Comprehension and Oral Narration. The Narrative Comprehension scale requires children to answer comprehension questions about three oral stories. The Oral Narration scale requires children to retell one oral story using no visual prompts and tell two oral stories, one while viewing a sequence of five pictures and another while viewing a single picture. For the TNL English (Gillam & Pearson, 2004), Hispanic children make up 12% of the normative sample and the measure has been validated for use with bilingual children (Gillam et al., 2013). Sensitivity and specificity for confirming the presence or absence of language disorder are 0.92 and 0.87, respectively. For the TNL Spanish (Gillam et al., 2010), preliminary data based on 216 children suggest alpha levels of 0.89 and 0.93 for the Narrative Comprehension and Oral Narration scales, respectively (Peña et al., 2020). Furthermore, data from a subset of 90 children showed that children with typical language development receive significantly higher raw scores on the TNL Spanish subtests (M = 8.6) than children with DLD (M = 4.4).

Index Measure

A subset of test items from the BESA-ME morphosyntax subtests was selected to serve as the index measure for evaluating the feasibility of ASR for DLD identification in the current study. The BESA-ME morphosyntax was selected as the index measure for two primary reasons. First, unlike the BESA-ME semantics subtest and TNL, all items on the BESA-ME morphosyntax subtest elicit expressive productions and are therefore candidates for automatic ASR scoring. Second, in an analysis of second and fourth grade Spanish-English bilinguals, Peña and colleagues (2020) demonstrated that the BESA-ME morphosyntax accounted for the most variance in discriminating between TD and DLD second graders (the age group of interest in the current study), over and above the BESA-ME semantics and TNL.

To ensure test items' usability for disorder identification purposes and with children with varying degrees of English language exposure, however, only BESA-ME morphosyntax items that met the following criteria were included in the index measure:

(a) an item discrimination index at or above 0.30 between TD and DLD children (Ebel & Frisbie, 1986),

(b) an item mean difference at or above 0.30 between TD and DLD children, and

(c) an item mean response value at or above 0.30 for at least two of three language exposure profiles (English-dominant [60% or more current English exposure], Spanish-dominant [40% or less current English exposure], or balanced [40 to 60% current English exposure]).

In addition, given that a subset of items represented individual word or phrase targets from sentence repetition test items, only sentences for which more than one target met the inclusionary criteria were included in the index measure. To identify the test items that met these inclusionary criteria, the item-level data of all children who completed the BESA-ME

morphosyntax subtests during the larger longitudinal study (Peña et al., 2010a) were analyzed. This sample included 283 children (TD = 237, DLD = 46) who completed the English morphosyntax subtest, contributing an average of 2.3 datapoints per English item, and 252 children (TD = 212, DLD = 40) who completed the Spanish morphosyntax subtest, also contributing an average of 2.3 datapoints per Spanish item. Children's classifications were determined using the DLD identification protocol of the original study, discussed above. At each time point, children's documented current language exposure from the BIOS was used to determine their language exposure profiles. Item discrimination indices were calculated using the following formula (Wood, 1960):

$$\frac{\#\ of\ TD\ children\ responding\ correctly - \#\ of\ DLD\ children\ responding\ correctly}{Total\ \#\ of\ TD\ children}$$

This produced a value ranging from -1.0 to 1.0, with a value of 1.0 indicating that 100% of participants in the TD group and 0% of participants in the DLD group responded correctly to that test item. Classification means were calculating by averaging all responses to an item from children who fell into one of the two classification groups. Finally, language exposure means were calculated by averaging all responses to an item from children who fell into one of the three language exposure profiles at the time the response was elicited. This procedure resulted in 34 English and 27 Spanish items (listed in Appendix A) from the original 102 English and 108 Spanish BESA-ME morphosyntax items. Discrimination indices ranged from 0.43 to 0.78. The average discrimination index for the English items was 0.67 (*SD* = 0.09), while the average discrimination index for the Spanish items was 0.61 (*SD* = 0.09).

Automatic Speech Recognition Application

The ASR application used to transcribe children's test responses for the current analyses was a researcher-coded Python program that employed version 1 of the Google Cloud non-

streaming REST speech-to-text API (Google, 2020). The REST API is a programmable algorithm developed by Google that asynchronously converts human speech to text across 125 languages and language variants. While it is commercially available in multipurpose consumer devices and software (e.g., Google Drive), it is also available for use in custom applications at a cost per minute basis. To include the REST API in a custom program, a JavaScript object notation (JSON) access token associated with a Google Cloud account is written into a developer's custom code using the programming language of choice (Python was used in the current study). The access token then allows the custom program to send audio data to the Google server, where it is converted to text and returned to the user. As there are multiple language and model options, the code is programmable for the target language and target type of model. In the current study, the *en-US* (United States English) and *es-US* (United States Spanish) "command and prompt" models, which Google specifies are suitable for analyzing short segments of speech (Google, 2020), were employed.

Several critical features of the API, reported by Google (2020), made it a suitable candidate software for the current study. First, it transcribes United States variants of English and Spanish, which were spoken by the children in this study. Second, the API uses "noise robustness," a feature that allows the software to handle noisy audio without requiring noise cancellation. This is beneficial given that the recordings used in this study were collected at children's schools and often contained background noise. Third, the API employs speech recognition models that have been pre-trained with millions of hours of speech data from both adults and children, increasing its potential accuracy on child speech and precluding the need to conduct independent model training. Finally, data-logging is opt-in only (i.e., the API does not log data by default), ensuring the privacy of children's audio files.

PROCEDURES

Data Collection

During the screening phase of the longitudinal study, children completed the BESOS in English and Spanish and parents and teachers completed the BIOS. Trained Spanish-English bilingual research assistants administered the BESOS to children individually at their schools. Generally, all subtests and languages of the BESOS were completed within a single one-hour testing session. If testing was cut short due to scheduling conflicts or if additional time was required, research assistants returned to the schools on a different day to complete testing. Children's responses were recorded on paper test forms that were later digitally scanned and uploaded to a secure file server. Parents completed the Home BIOS via a phone interview with a bilingual research assistant in the parent's preferred language. Teachers completed the School BIOS in person, either filling out the form individually or with guidance from a research assistant. All BIOS responses were recorded on paper test forms that were later digitally scanned and uploaded to a secure file server.

During the testing phase of the longitudinal study, which began one year after the screening phase, children completed a battery of language and cognition measures once per year for up to four years. The battery included the BESA-ME, which is of interest to the current analysis. Children also received the Bilingual English Spanish Assessment, Universal Nonverbal Intelligence Test, Test of Narrative Language, Expressive One-Word Picture Vocabulary Test (Brownell, 2000; Brownell, 2001), and experimental measures of nonword repetition, semantic blocking, grammatical priming, and tongue twister production. These are not of interest to the current analysis. Bilingual research assistants administered all test measures to children

individually at their schools. While some children were tested in empty rooms or areas of the school, others were tested in the same area as other children or adults. Sufficient physical distance was placed between children such that evaluators' ability to administer and score the measures was not impacted and children's responses to test items were not influenced by their peers. However, this setup occasionally caused background noise to occur during testing. Testing was completed over three to six sessions that were 30 minutes to an hour in length. Children's responses were manually recorded on paper test forms and audio recorded using Zoom H2n Handheld SD Recorders in .mp3 320kbps acg2 (for speech) mode. The scanned paper test forms and digital audio recordings were later uploaded to a secure file server. Parents and teachers completed the BIOS and ITALK following the same BIOS procedure from the screening phase.

Data Analysis

Audio Recording Processing

There were two existing audio recordings per child, one from each of their BESA-ME morphosyntax testing sessions (English and Spanish). Children's responses to the target test items were extracted from the longer audio recordings using Audacity Version 2.3.2 (Audacity Team, 2020), yielding an individual audio recording for each test item response. This procedure was conducted because the longer audio recordings contained examiners' prompts and in order to simulate the length of the responses that would be obtained if children were completing the test with a conversational agent employing ASR. The segmented audio files were saved in the .wav file format at the original 16,000 Hz sampling rate and mono signal.

Automatic Speech Recognition Transcription

To convert children's audio files to ASR-transcribed item responses, the researcher-coded ASR speech-to-text Python program conducted the following processes: (a) extract audio

recording file from a given local directory, (b) scan the file name for the target language (English or Spanish), (c) convert the entire audio response to text using the target language speech-to-text transcription model, generating up to four transcription alternatives, and (d) output the transcription alternatives to a .csv file, with one row representing one audio file.

Scoring

Children's original scores on the test items, scored by human evaluators during testing, were drawn from the existing dataset. Scoring reliability was performed on 10% of the samples from the original longitudinal study and yielded an average 99.8% interrater reliability, ensuring the reliability of the human evaluators. To determine children's ASR-transcribed scores, for each item, children's transcripts were programmatically compared to the target responses using an R script. Children were assigned a score of 1 or 0 on an item depending on whether the ASR-transcribed response across any of the child's four transcription alternatives included or did not include the target response for that item, respectively. All possible target responses were derived from the BESA-ME morphosyntax record protocols.

Chapter 3: Results

SCORING AGREEMENT

The first aim of this study was to determine the item-by-item agreement between children's human-scored test scores and their ASR test scores on the subset of BESA-ME morphosyntax items and to explore whether agreement varied due to child-level or item-level factors. There were 5,124 total item-level responses. To estimate agreement while accounting for the possibility of chance agreement, I calculated the Cohen's kappa coefficient between the human and ASR scores across disorder statuses, language exposure groups, and at the item level.

Disorder Status

To determine whether scoring agreement varied by disorder status, I calculated the Cohen's kappa coefficient between the human and ASR scores by disorder status, test language, and item type. The results of these calculations are shown in Table 2.

Table 2. *Scoring Agreement Between the Human and ASR Scores of the BESA-ME Morphosyntax Subtest by Disorder Status*

	TD	DLD	Total
Overall	0.52 [0.50-0.55]	0.36 [0.31-0.41]	0.54 [0.52-0.56]
English	0.45 [0.42-0.49]	0.27 [0.20-0.35]	0.47 [0.44-0.50]
Cloze	0.44 [0.39-0.49]	0.25 [0.14-0.37]	0.46 [0.42-0.51]
Sentence Repetition – Targets	0.44 [0.38-0.49]	0.28 [0.17-0.38]	0.47 [0.42-0.51]
Sentence Repetition – Verbatim	0.28 [0.11-0.44]	0.00 [0.00-0.00]	0.28 [0.11-0.44]
Spanish	0.61 [0.57-0.65]	0.44 [0.36-0.51]	0.62 [0.59-0.65]
Cloze	0.55 [0.46-0.64]	0.55 [0.39-0.70]	0.60 [0.53-0.67]
Sentence Repetition – Targets	0.62 [0.58-0.67]	0.40 [0.31-0.49]	0.62 [0.58-0.66]
Sentence Repetition – Verbatim	0.52 [0.38-0.66]	0.00 [0.00-0.00]	0.54 [0.41-0.67]

Note. Values outside the brackets represent the Cohen's kappa coefficients. Values inside the brackets represent the lower and upper 95% confidence intervals. Per Cohen (1960), kappa coefficient values of ≤ 0 = no agreement (light red), 0.01–0.20 = none to slight (not pictured), 0.21–0.40 = fair (light green), 0.41– 0.60 = moderate (medium green), 0.61–0.80 = substantial (dark green), and 0.81–1.00 = almost perfect agreement (not pictured). TD = typical language development, DLD = developmental language disorder.

The overall item-level agreement across classifications and test languages was 0.54, indicating moderate overall scoring agreement between the human and ASR scores. However, there were variations in agreement between test languages, classifications, and item types. To determine whether these variations were substantial, I evaluated the overlap between the 95% confidence intervals of the coefficients. There was no overlap in the overall English and Spanish confidence intervals, indicating that the Spanish subtest yielded higher agreement ($k = 0.62$) than the English subtest ($k = 0.47$). There was also no overlap in the TD and DLD confidence intervals overall, by test language, or by test item type (except for Spanish cloze), indicating that responses by children with TD generally yielded higher agreement (overall $k = 0.52$) than responses by children with DLD (overall $k = 0.36$). All agreement coefficients for the TD group were in the moderate to substantial range ($k = 0.44$ to $k = 0.62$) except for agreement on Spanish SR verbatim items, which was fair ($k = 0.28$). All agreement coefficients for the DLD group were in the fair to moderate range ($k = 0.25$ to $k = 0.55$) except for agreement on the Spanish and English SR verbatim items, which yielded no agreement ($k = 0$ for both). This distinct level of agreement between the TD and DLD groups poses a potential risk to the classification accuracy of the ASR measure. The results also highlight ASR's relatively poorer scoring accuracy on SR verbatim test items.

Language Exposure Group

To determine whether scoring agreement varied by language exposure group, I calculated the Cohen's kappa coefficient between the human and ASR scores by exposure group, test language, and item type. The results of these calculations are shown in Table 3.

Table 3. *Scoring Agreement Between the Human and ASR Scores of the BESA-ME Morphosyntax Subtest by Exposure Group*

	SD	B	ED
Overall	0.53 [0.49-0.56]	0.56 [0.53-0.59]	0.52 [0.44-0.60]
English	0.46 [0.42-0.51]	0.47 [0.42-0.51]	0.46 [0.36-0.57]
Cloze	0.44 [0.38-0.51]	0.44 [0.38-0.51]	0.54 [0.39-0.69]
Sentence Repetition – Targets	0.47 [0.40-0.53]	0.45 [0.39-0.52]	0.38 [0.22-0.54]
Sentence Repetition – Verbatim	0.32 [0.02-0.63]	0.31 [0.09-0.53]	0.00 [0.00-0.00]
Spanish	0.56 [0.51-0.61]	0.68 [0.64-0.72]	0.46 [0.31-0.61]
Cloze	0.54 [0.43-0.65]	0.64 [0.54-0.74]	0.53 [0.20-0.86]
Sentence Repetition – Targets	0.55 [0.50-0.61]	0.70 [0.65-0.75]	0.44 [0.27-0.61]
Sentence Repetition – Verbatim	0.49 [0.30-0.67]	0.59 [0.40-0.77]	0.00 [0.00-0.00]

Note. Values outside the brackets represent the Cohen's kappa coefficients. Values inside the brackets represent the lower and upper 95% confidence intervals. Per Cohen (1960), kappa coefficient values of ≤ 0 = no agreement (light red), 0.01–0.20 = none to slight (not pictured), 0.21–0.40 = fair (light green), 0.41– 0.60 = moderate (medium green), 0.61–0.80 = substantial (dark green), and 0.81–1.00 = almost perfect agreement (not pictured). SD = Spanish-dominant (at least 60% current Spanish exposure), B = balanced (between 40-60% current English and Spanish exposure), ED = English-dominant (at least 60% current English exposure).

Agreement ranged from moderate to substantial (k = 0.44 to k = 0.70) and the 95% confidence intervals overlapped across all three language exposure groups for all but five coefficients. The English SR verbatim agreement was fair for the SD (k = 0.32) and B (k = 0.31) children, but these confidence intervals overlapped with the other two item types, indicating little distinction between the three English item types for these groups. Similarly, the ED group had fair agreement (k = 0.38) on the English SR target items, but this confidence interval overlapped with that of the English cloze items, indicating that the two item types were similar for this group. However, for the ED group, there was no agreement (k = 0) on the English and Spanish SR verbatim items. Overall, these findings indicated that there were generally similar levels of agreement between the three exposure groups across test languages and item types, demonstrating minimal differential effects of language exposure on ASR's scoring of test items. The findings also further disambiguated the unique difficulty of the SR verbatim scoring,

showing that ASR agreed on SR verbatim items least when children were English-dominant (regardless of test language).

Item-Level

To identify scoring agreement by item and target form, the Cohen's kappa coefficient was individually calculated for each test item (listed in Appendix A). Item-level agreement ranged from slight (with a minimum of 0.08 on an English SR verbatim item) to almost perfect (with a maximum of 0.85 on a Spanish relative clause item), indicating substantial variation in agreement across test items. Of the 61 test items, three yielded agreement above 0.80 (Spanish = 2, English = 1), 18 yielded agreement between 0.60-0.80 (Spanish = 13, English = 5), 25 yielded agreement between 0.40-0.60 (Spanish = 9, English = 16), 12 yielded agreement between 0.20-0.40 (Spanish = 2, English = 10), and three yielded agreement below 0.20 (Spanish = 1, English = 2). These results demonstrated that most test items yielded at least moderate (i.e., at or above 0.40) agreement, but highlighted the subset of items with poorer agreement that may be candidates for omission in later analyses.

Finally, a descriptive analysis of agreement by morphosyntactic form was conducted to provide further information about the best candidate item types for ASR. English passive (e.g., is pulled, is going), progressive (e.g., go**ing**), and past tense (e.g., dropp**ed**) items yielded generally poorer agreement (k = 0.16 to k = 0.34). English items capturing the third person singular, possessive, or plural -s (i.e., with an "-s" word ending) yielded higher but mixed agreement (k = 0.34 to k = 0.73). English infinitive ("**to get**"), question inversion (e.g., "**can she**…"), and plural ("**boxes**") items yielded the highest agreement (k = 0.60 to k = 0.73). In Spanish, all but two forms (indirect object clitic ["**les** (dice)"] and imperfect ["**montaban**"]) had agreement above 0.50. The forms with the highest agreement were preterites ("**vio**", "**pidio**"), negatives

("**ningun**"), and relative clauses ("**que le gusta**") (k = 0.73 to k = 0.85). These results demonstrated the generally higher agreement among Spanish test items compared to English test items and disambiguated the morphosyntactic forms with relatively lower and higher mean agreement coefficients on this measure. It is important to note, however, that because a relatively small subset of items was used in this study, most morphosyntactic forms were represented by only one or two test items, thus limiting the extent to which the findings about target forms can be generalized.

Table 4. *Average Scoring Agreement by Language and Morphosyntactic Form*

Language	Form	*n*	Mean *k*	Targets
English	Passive	2	0.16	*Is pulled, is lifted*
	SR Verbatim	2	0.23	*Sentence repeated verbatim*
	Present Progressive	1	0.30	*Is going*
	Past Tense (Regular)	2	0.34	*Planted, dropped*
	Third Person Singular	3	0.34	*Reads, eats, wears*
	Past Tense (Irregular)	2	0.40	*Took, went*
	Conjunction	1	0.43	*If*
	Possessive 's	3	0.45	*Chef's, clown's, cowboy's*
	Demonstrative	1	0.49	*Those (fish)*
	Relative Pronoun	4	0.49	*That*
	Pronoun	2	0.50	*She*
	Preposition	2	0.51	*In, for*
	Auxiliary	1	0.56	*Had*
	Infinitive	1	0.60	*To get*
	Question Inversion	6	0.60	*Can she, where is, etc.*
	Plural	1	0.73	*Boxes*
Spanish	Indirect Object Clitic	1	0.11	*Les (dice)*
	Imperfect	1	0.28	*Montaban (caballos)*
	Conditional	1	0.51	*Montaria*
	Preposition	5	0.52	*A, para, del, de*
	SR Verbatim	2	0.54	*Sentence repeated verbatim*
	Subjunctive	4	0.58	*Que le de, que vayan*
	Conjunction	1	0.60	*Aunque*
	Relative Pronoun	3	0.63	*Que, cuando*
	Possessive Article	1	0.65	*Su*
	Plural Adjective	1	0.68	*Felices*

Table 4. (continued)

Adverb	1	0.69	*Siempre*
Preterite	3	0.73	*Vio, recibio, pidio*
Negative	2	0.78	*Ningun, ninguna*
Relative Clause	1	0.85	*Que le gusta*

Note. n = number of items, k = Cohen's kappa coefficient

CLASSIFICATION ACCURACY

The second aim of this study was to determine the classification accuracy of children's ASR test scores, i.e., the extent to which ASR scores accurately grouped children into the DLD and TD groups. Children's existing disorder classifications were used as the reference for examining classification accuracy. Classification analyses were conducted in two stages. First, because this study analyzed a subset of items from the BESA-ME Morphosyntax, the classification accuracy of the human-scored item subset was established (i.e., the human-scored condition). Second, the classification accuracy of the ASR-transcribed and programmatically-scored item subset was determined (i.e., the ASR condition).

In keeping with prior research, children's best language scores (the higher percentage correct score of the two languages) were entered into all classification analyses. To disambiguate whether scores from different elicitation methods yielded superior classification accuracies, scoring was divided into seven methods. The first four methods combined scores across scoring types: total score, cloze + SR targets, cloze + SR verbatim, and SR targets + verbatim (i.e., SR total). The remaining three methods used the individual scoring types: cloze only, SR targets only, and SR verbatim only. All classification analyses were conducted using receiver operating characteristic (ROC) curve analyses. The ROC curve analyses identified the thresholds for each of the scoring methods (i.e., the percentage correct scores that maximized sensitivity and

specificity, serving as the optimal cut point for discriminating between children with and without DLD) and the classification metrics associated with each threshold.

Human-Scored Classification Accuracy

The results of the ROC curve analyses of children's human-scored items are shown in Table 5.

Table 5. *Receiver Operating Characteristic Curves on Children's Human-Scored Test Scores (DLD = 25, TD = 59)*

	Total	Cloze + SR Targets	Cloze + SR Verbatim	SR Total	Cloze Only	SR Targets Only	SR Verbatim Only
Threshold	54	61	32	58	61	62	25
Accuracy	92	91	86	87	89	89	81
Sensitivity	88	92	84	84	84	84	80
Specificity	93	90	86	88	92	92	81
True Positives	22	23	21	21	21	21	20
False Negatives	3	2	4	4	4	4	5
True Negatives	55	53	51	52	54	54	48
False Positives	4	6	8	7	5	5	11
Positive Likelihood Ratio	12.98	9.05	6.20	7.08	9.91	9.91	4.29
Negative Likelihood Ratio	0.13	0.09	0.19	0.18	0.17	0.17	0.25

Note. DLD = developmental language disorder [positive cases], TD = typical language development [negative cases], SR = sentence repetition

At the identified thresholds, all seven human-scored methods yielded adequate to good sensitivity (between 80% and 92%) and specificity (between 81% and 93%) for identifying DLD, per Plante and Vance (1994), with a maximum of 92% of cases correctly classified (total scoring). The total score yielded the highest positive likelihood ratio (12.98), indicating that this scoring method was associated with the highest likelihood that a positive result (i.e., a DLD case) was true (Dollaghan, 2007). The cloze + SR target score yielded the lowest negative likelihood ratio (0.09), indicating that this scoring method was associated with the highest likelihood that a negative result (i.e., a TD case) was true. These results indicated that, in the human-scored

condition, all seven scoring methods could be used to classify children with and without DLD. As such, the human-scored condition served as a robust baseline from which to analyze the ASR condition.

ASR Classification Accuracy

To determine the relationships between the human- and ASR-scored items, the item discrimination indices of the ASR-scored items were calculated. These, shown in Appendix A, were significantly and positively correlated with the human-scored item discrimination indices, $r = 0.45$, $p = 0.01$, indicating their concurrent validity, but were lower overall, ranging from 0.03 to 0.63. The average discrimination index for the English items was 0.30 ($SD = 0.12$), while the average discrimination index for the Spanish items was 0.35 ($SD = 0.13$). Paired t-tests comparing the human and ASR discrimination indices confirmed that the ASR indices were significantly lower than the human-scored indices, $t = -9.689$, $p < 0.001$, with a mean difference of -0.14. Furthermore, discrimination indices were significantly and positively correlated with items' Cohen's kappa coefficients, $r = 0.45$, $p = 0.0003$, indicating that human-ASR item agreement was positively associated with the ASR discrimination index of that test item. These findings suggested that the ASR classification analyses would yield similar accuracies to the human-scored classification analyses, but that some items with lower scoring accuracies and/or discrimination indices in the ASR-scored condition may negatively impact the ASR results.

The results of the ROC curve analyses of children's ASR-transcribed and programmatically scored test scores are shown in Table 6.

Table 6. *Receiver Operating Characteristic Curves on Children's ASR Test Scores (Original Item Set) by Scoring Method (DLD = 25, TD = 59)*

	Total	Cloze + SR Targets	Cloze + SR Verbatim	SR Total	Cloze Only	SR Targets Only	SR Verbatim Only
Threshold	39	40	19	37	37	27	25
Accuracy	88	88	86	87	89	92	65
Sensitivity	88	88	88	84	72	72	100
Specificity	88	88	85	88	97	100	51
True Positives	22	22	22	21	18	18	25
False Negatives	3	3	3	4	7	7	0
True Negatives	52	52	50	52	57	59	30
False Positives	7	7	9	7	2	0	29
Positive Likelihood Ratio	7.42	7.42	5.77	7.08	21.24	Inf	2.03
Negative Likelihood Ratio	0.14	0.14	0.14	0.18	0.29	0.28	0

Note. DLD = developmental language disorder [positive cases], TD = typical language development [negative cases], SR = sentence repetition

At the identified thresholds, all four combined ASR scoring methods yielded adequate sensitivities (85% to 88%) and specificities (between 85% and 88%) for identifying DLD in the ASR condition, with a maximum of 88% of cases correctly classified (total and cloze + SR target scoring). The individual scoring methods showed mixed sensitivity and specificity. Specifically, SR target scoring and cloze only scoring had extremely high specificities (97% and 100%, respectively) at the cost of sensitivity, which was inadequate at 72% for both methods. Conversely, SR verbatim scoring had extremely high sensitivity (100%) at the cost of specificity, which was at chance (51%). These mixed results also led to the individual methods yielding the highest positive likelihood (cloze = 21.24; SR targets = Inf) and lowest negative likelihood (SR verbatim = 0) ratios, but the costs associated with these individual scoring methods made them inadequate for classification purposes in the ASR condition. These results demonstrated that only ASR scores from the four combined scoring methods adequately classified children with and without DLD.

Compared to the human-scored classification analyses, specificity was slightly lower for the ASR condition on three of the four combined scoring methods (ranging from 1% to 5% lower), while specificity of the SR total score was the same across the human-scored and ASR conditions (88%). Sensitivity did not substantially differ. The ASR condition yielded higher sensitivity than the human-scored condition when cloze plus SR target scoring was used (88% compared to 84%) and the same sensitivity as the human-scored items when total scoring was used (88%). These results indicated that ASR was slightly less accurate at identifying TD cases, generally yielding a higher number of false positives, but identified DLD cases comparably to the human-scored condition. Additionally, of note was that the thresholds for the ASR scores were 12% to 35% lower than the thresholds in the human-scored condition across six of the seven scoring methods (except for SR verbatim method, which had an identical threshold of 25% for the ASR and human-scored items). This indicated that, generally, a lower percentage correct score discriminated between children with and without DLD when ASR was used to transcribe and score their test responses.

Follow-Up Classification Analyses

To explore whether the classification accuracy of the ASR scores could be improved from adequate to good, an item selection procedure was conducted, test items were omitted, and classification analyses were repeated on a new shorter item set. Only ASR items that were likely to increase classification accuracy were retained in this shorter item set. Specifically, I retained only ASR items with a discrimination index at or above 0.3, a DLD-TD mean difference at or above 0.3, and a mean score at or above 0.3 for two of the three language exposure groups (the same item selection criteria used to construct the index measure). This yielded 15 English items, of which six were cloze items and nine were SR target items (from four distinct sentences), and

17 Spanish items, of which three were cloze items and 14 were SR target items (from six distinct sentences). These items are identified in Appendix A. None of the retained items were SR verbatim items.

To determine the classification accuracy of this shorter set, I conducted ROC curve analyses on the best language percentage correct score using each of the three possible scoring methods: total, cloze only, and SR targets only. The results are shown in Table 7.

Table 7. *Receiver Operating Characteristic Curves on Children's ASR Test Scores (Shorter Item Set) by Scoring Method (DLD = 25, TD = 59)*

	Total	Cloze Only	SR Targets Only
Threshold	38	42	53
Accuracy	94	89	86
Sensitivity	84	84	96
Specificity	98	92	81
True Positives	21	21	24
False Negatives	4	4	1
True Negatives	58	54	48
False Positives	1	5	11
Positive Likelihood Ratio	49.56	9.91	5.15
Negative Likelihood Ratio	0.16	0.17	0.05

Note. DLD = developmental language disorder [positive cases], TD = typical language development [negative cases], SR = sentence repetition

All three scoring methods yielded adequate to good sensitivity (84% to 96%) and specificity (81% to 98%) with a maximum of 94% of cases correctly classified (total scoring), indicating that the shorter set of items could be used for DLD identification and that there were some improvements to the classification metrics. Classification accuracy was higher overall for the shorter ASR set (mean across scoring methods = 89.7%) compared to the longer ASR set (mean across scoring methods = 85.0%), but there were some costs associated with this. Specifically, for the total score (which combined the cloze and SR target items), specificity improved at a small cost to sensitivity. Specificity improved 10% (from 88% to 98%), but

sensitivity dropped 4% (from 88% to 84%). A reverse pattern of results was observed for the individual cloze and SR target scores, with sensitivity improving at a cost to specificity. Compared to the longer set, the sensitivity of the cloze only score on the shorter set improved 12% (from 72% to 84%), while specificity dropped 5% (from 97% to 92%). Similarly, the sensitivity of the SR target score improved 24% (from 72% on the longer set to 96% on the shorter set), while specificity dropped 19% (from 100% to 81%).

Notably, the shorter set total score yielded the highest positive likelihood ratio (49.56) among all classification analyses conducted in this study with at least adequate sensitivity and specificity, indicating that it was the method associated with the highest likelihood that a positive result (i.e., a DLD case) was true. Similarly, the shorter set SR target score yielded the lowest negative likelihood ratio (0.05) among all classification analyses conducted in this study with adequate sensitivity and specificity, indicating that it was the method associated with the highest likelihood that a negative result (i.e., a TD case) was true. These results demonstrated that reducing the original item set improved the overall classification accuracy of the ASR scores, leading to more accurate identification of DLD and TD. Moreover, unlike the longer ASR item set, the shorter ASR item set yielded very high values on some classification metrics while maintaining adequate values on all other classification metrics.

Chapter 4: Discussion

This study presents preliminary evidence for the technical feasibility of ASR as a bilingual expressive language assessment tool. The dual-language morphosyntax assessment responses of Spanish-English bilingual second graders with and without confirmed diagnoses of DLD were used to develop a bilingual English-Spanish index measure with high classification accuracy when scored by a human examiner. Children's audio-recorded responses to the items on this measure were transcribed by a researcher-developed ASR application and programmatically scored. The ASR measure achieved moderate item-by-item scoring agreement with the human-scored measure overall and there were minimal variations in agreement between language exposure groups. There were differences in agreement across other factors, including TD children yielding higher agreement than DLD children, Spanish test items yielding higher agreement than English test items, and cloze and sentence repetition target items yielding higher agreement than sentence repetition verbatim items. Furthermore, item-by-item agreement was significantly associated with item discrimination indices, indicating that higher agreement would yield improved classification.

Despite the variability in scoring agreement between the human- and ASR-scored measures, their classification accuracies differed by just four percentage points (92% and 88% of cases classified correctly, respectively, using the total best-language percentage correct), with the ASR measure yielding the same sensitivity but lower specificity. When the ASR-scored test items were further narrowed by retaining only those with adequate or higher discrimination in the ASR condition, accuracy rose to 94% of cases classified correctly using the total best-language percentage correct. This increase was exclusive to improved specificity. These findings demonstrated that an identical ASR adaptation of an existing expressive morphosyntax measure

achieved the same identification of DLD but slightly lower identification of TD, but that the ASR test item subset could be manipulated to increase specific classification metrics.

The Path Toward Automatic Assessment

The current findings suggest that children's expressive language skills (i.e., their verbal responses) in more than one language can be automatically evaluated. Specifically, I extend the work of de Villiers and colleagues (2021) and Pratt and colleagues (in review), who demonstrated the validity of automatic receptive language tasks (i.e., listening to prompts and clicking/tapping the correct responses) in two languages. In this study, children's verbal responses to English and Spanish expressive test items were successfully automatically transcribed and scored, yielding fair to good classification accuracy. Together with the previous research, these findings provide evidence for the feasibility of assessment instruments that automatically administer and score language tasks across both the expressive and receptive modalities for DLD identification purposes.

There are two key contributions to the field of child language assessment associated with this outcome. First and more broadly, this study shows that ASR scoring of children's morphosyntax skills is viable within languages and item response types, suggesting that a range of measures can employ this technology. That is, scoring agreement was at least moderate for both the English and Spanish items and both the cloze and sentence repetition items. These results suggest that single-language English or Spanish measures and/or measures employing one or both response elicitation methods can be scored using ASR. This is an important contribution in that it supports a path toward the broad adoption of automatic DLD assessment instruments, which has the potential to improve the efficiency and accuracy of language assessment practices for all children, not just those who are bilingual.

A second key contribution and one more specific to the current aims is that this study supports a novel method for SLPs to validly and reliably assess languages they do not speak. Automatic dual-language assessment may reduce the reliance on bilingual SLPs and interpreters, who can be inaccessible or who may not have the resources to conduct such assessments (e.g., Arias & Friberg, 2017; Saenz & Langdon, 2019). This would allow all SLPs to collect information about both languages of a bilingual child, a practice necessary for accurate DLD diagnosis, reducing the risk of misdiagnosis in this population that can lead to endemic inequities in educational access.

An important consideration is that both mentioned contributions are conditional upon SLPs' adoption of such tools, but the current evidence suggests that it is a matter of when and not if automatic assessments are likely to be adopted. There is emerging evidence that automatic expressive language tasks are practically feasible, with children as young as three successfully engaging in these tasks for as long as 30 minutes (Xu et al., 2021; Yeung et al., 2019). Additionally, while this study did not explore SLPs' attitudes about the adoption of ASR for language assessment, a significant factor predicting SLPs' use of clinical technologies is whether the technologies enable them to accomplish tasks more quickly and effectively (Albudoor & Peña, 2021; Boster & McCarthy, 2018). Finally, similar tools are prevalent in K-12 education, suggesting that their adoption and implementation is likely. All three of the most common K-12 English language proficiency measures in the United States—ACCESS, ELPA21, and ELPAC—are electronically administered and partially automatically scored on desktop or laptop computers (Kim et al., 2020). These tools evaluate children's listening, speaking, reading, and writing skills using tasks very similar to those employed by DLD identification instruments and have been adopted by 49 of the 50 U.S. states. While the mentioned proficiency measures do not

yet automatically score children's verbal or written expressive responses, test development companies are now trialing automated speech scoring systems for child language proficiency measures such as the Test of English as a Foreign Language (TOEFL) Junior (Evanini et al., 2020). Together, these findings indicate that more widespread implementation of automatic language assessments employing ASR is likely to occur and that SLPs are likely to adopt such technologies if they are available and effective, further confirming the importance of providing empirical support for their use in DLD identification.

Identifying What Works and What Does Not in Automatic Language Assessment

This study provisionally disambiguates what works from what does not in the automatic assessment of English and Spanish skills for the purposes of DLD identification, providing four critical considerations for automatic assessment.

First, this study established that there was some cost to TD-DLD discrimination associated with ASR scoring, but that the original (conservative) item selection procedure prevented the ASR classification accuracy from dropping substantially. The individual discrimination indices of the test items fell significantly in the ASR condition compared to the human-scored condition, but the overall classification accuracy of ASR was only four percentage points lower and was able to be increased in follow-up analyses. Even when test items were further restricted in the follow-up analyses, there remained enough items to re-conduct the classification analyses and yield good classification accuracy. These findings suggested that it was important to begin with a larger but highly robust item subset as the index measure, as some items in the ASR condition yielded low enough discrimination indices to be candidates for elimination.

Second, the current findings highlighted how ASR classification accuracy could vary in different directions from human-scored classification accuracy. In the first set of ASR analyses, in which all items were tested, classification accuracy only dropped in specificity (i.e., the measure's ability to identify true negative [TD] cases) but not in sensitivity (i.e., the measure's ability to identify true positive [DLD] cases). In other words, the ASR falsely flagged more children as DLD, indicating that, all things being equal, children are more likely to fail an ASR-scored measure compared to a human-scored measure. However, the reverse pattern was observed when the ASR measure was modified to include only items with adequate discrimination indices, with the ASR falsely flagging more children as TD. This demonstrated that ASR sensitivity and specificity did not vary in a single direction. This is a broadly positive finding in that it confirms that ASR is not consistently poorer at identifying a single class of cases. That is, there is no bias toward classifying children as TD or DLD. Together with the results that ASR did not bias a specific language exposure group, these are positive indicators of ASR's robustness to child- and group-level variations. Furthermore, the mixed results demonstrate that it is possible to modify an ASR item subset to achieve the specific classification metrics necessary for the purposes of the test. For example, an ASR test developed for screening purposes may prefer a higher false positive than false negative rate, to ensure that children with DLD are not overlooked. Conversely, an ASR test developed for diagnostic purposes may prefer to have the highest likelihood that a positive result is true, to increase confidence that DLD diagnoses are accurate.

A third consideration was that certain item elicitation types may or may not be good candidates for ASR scoring at present. As mentioned, ASR generally reliably scored test items that required children to complete sentences (i.e., cloze) or that confirmed whether target words

or phrases in sentences were repeated. However, items requiring children to repeat sentences verbatim yielded poor scoring agreement and classification accuracy. These findings indicated that ASR is not yet sensitive enough to reliably confirm whether every single word in a given sentence is repeated by a child. Although the present study did not compare the word-, phrase-, or sentence-level accuracy of ASR, this finding is unsurprising given the ASR accuracy rates reported by other researchers. For example, Hair and colleagues (2019) reported an ASR accuracy rate of 90% on the word-level responses of children with speech disorders. While high, a 90% rate suggests that one in ten words in a given sentence will be incorrectly processed by ASR, indicating that verbatim sentence repetition scoring is likely too stringent for this scoring method at present. It is possible that less stringent criteria for sentence repetition scoring (e.g., 80% of targets detected) may yield higher ASR-human scoring agreement, but additional analyses are necessary to establish this.

Fourth and finally, the current results provided preliminary evidence about the morphosyntactic targets that may or may not be good candidates for ASR scoring. Broadly, Spanish test items yielded higher scoring agreement than English test items. When agreement by morphosyntactic form was descriptively examined, all but two Spanish forms (indirect object clitic and imperfect) and all but five English forms (passive, present progressive, regular past tense, irregular paste tense, and third person singular) yielded lower than moderate agreement, confirming the bias toward Spanish forms but demonstrating that most forms sampled in this study were adequately scored by the ASR. Notably, nearly all English forms that yielded lower agreement required the processing of bound morphemes (e.g., kick**ed**). Conversely, nearly all Spanish forms were represented by free morphemes (e.g., **ningun**). It is possible that the lower perceptual salience of the bound morphemes (Goldschneider & DeKeyser, 2001) made it

difficult for ASR to detect them. That is, their potentially shorter duration, lower volume, and/or lack of a segment boundary made it difficult for the ASR to identify whether they had been produced. However, this study also provided evidence against this explanation. The possessive **'s**, regular plural, and plural adjective forms, all of which are also bound, were at least moderately successfully scored by ASR. These findings suggest that a more discrete disambiguation of the morphosyntactic forms and their features is necessary to further explain why certain forms are less successfully scored by the ASR than others.

Limitations and Future Directions

There were two primary limitations of this study that represent directions for future research. First, this study was a secondary analysis of children's existing assessments, which were originally conducted by bilingual examiners. As such, the findings do not establish the practical feasibility of an ASR tool that uses automatic administration and is employed during live dual-language assessment. There is some evidence to suggest that results may differ due to technical challenges arising during live administration (e.g., Yeung et al., 2019). Furthermore, if the current measure is to replace the need for bilingual personnel, it is necessary to establish whether children can complete the tasks without the supervision of a bilingual practitioner. Therefore, a critical next step for this work is to test automatically administered and scored dual-language tasks during live assessment to determine the extent to which the current results hold and to establish whether it is possible for a practitioner who does not speak the test language to oversee the procedure.

A third and final limitation of this study was that it evaluated a small subset of items from a single linguistic domain, morphosyntax. Although the reduction of test items was conducted to maximize classification accuracy, it resulted in an index measure that sampled only one to two

exemplars of most morphosyntactic forms. This limited the extent to which conclusions could be made about the morphosyntactic forms and targets that may be better or worse candidates for ASR assessment. Future work sampling multiple exemplars of each form would provide more evidence for the words and phrases that are best analyzed by ASR, further informing research on automatic expressive language assessment. Finally, looking beyond morphosyntax, children with DLD demonstrate deficits in other language areas (e.g., semantics, narratives) that may also be candidates for ASR assessment and can provide examiners with a broader picture of a child's language skills. Previous work (Albudoor et al., 2019) provided early evidence that a broader subset of test items including probes across linguistic domains could achieve adequate scoring agreement with ASR. However, more research is needed to determine whether a cross-domain ASR measure can yield adequate classification accuracy and/or provide more comprehensive information about a child's language skills.

CONCLUSION

This study provides preliminary support for the technical feasibility of ASR for processing bilingual children's expressive dual-language assessment responses. While the current results are limited in their scope, they represent a proof of concept for the use of ASR in automatic language assessment instruments that test more than one language. Given the barriers to dual-language assessment that have contributed to disproportionate DLD misdiagnosis among bilinguals, this study presents critical evidence toward the expansion of access to more accurate assessment and intervention practices for this population.

Appendix A

Table 8. *Item-Level Information and Statistics, Sorted by Language and ASR Discrimination Index*

Test Language	Item Name & Number	Elicitation Type	Target Form	Human-ASR Agreement (Cohen's kappa)	Human Discrimination Index	ASR Discrimination Index	Included in Shorter ASR Item Set
English	eng_ms_cl_questinv_150	Cloze	Question Inversion	0.69	0.65	0.59	Yes
English	eng_ms_sr_inf_176	SR Target	Infinitive	0.60	0.71	0.59	Yes
English	eng_ms_sr_prep_171	SR Target	Preposition	0.48	0.71	0.59	Yes
English	eng_ms_sr_rel_168	SR Target	Relative Pronoun	0.70	0.71	0.56	Yes
English	eng_ms_cl_questinv_145	Cloze	Question Inversion	0.57	0.64	0.51	Yes
English	eng_ms_sr_pro_277	SR Target	Pronoun	0.55	0.75	0.49	Yes
English	eng_ms_sr_prog_175	SR Target	Present Progressive	0.30	0.78	0.49	Yes
English	eng_ms_cl_questinv_144	Cloze	Question Inversion	0.80	0.61	0.47	Yes
English	eng_ms_sr_aux_278	SR Target	Auxiliary	0.56	0.68	0.46	Yes
English	eng_ms_sr_conj_178	SR Target	Conjunction	0.43	0.77	0.46	Yes
English	eng_ms_sr_prep_204	SR Target	Preposition	0.54	0.70	0.46	Yes
English	eng_ms_sr_questinv_200	SR Target	Question Inversion	0.33	0.77	0.46	Yes
English	eng_ms_cl_poss_120	Cloze	Possessive 's	0.49	0.63	0.44	Yes
English	eng_ms_cl_questinv_146	Cloze	Question Inversion	0.68	0.60	0.42	Yes
English	eng_ms_sr_pro_169	SR Target	Pronoun	0.46	0.75	0.42	No
English	eng_ms_cl_questinv_148	Cloze	Question Inversion	0.54	0.59	0.41	Yes
English	eng_ms_cl_rel_137	Cloze	Relative Pronoun	0.50	0.69	0.39	No
English	eng_ms_cl_rel_136	Cloze	Relative Pronoun	0.52	0.70	0.34	No
English	eng_ms_cl_poss_003	Cloze	Possessive 's	0.46	0.69	0.32	No
English	eng_ms_cl_plural_027	Cloze	Plural	0.73	0.63	0.31	No
English	eng_ms_cl_rel_134	Cloze	Relative Pronoun	0.26	0.43	0.31	No
English	eng_ms_sr_pastirreg_074	SR Target	Past Tense (Irregular)	0.30	0.75	0.31	No
English	eng_ms_cl_passive_050	Cloze	Passive	0.23	0.75	0.29	No
English	eng_ms_sr_prorel_205	SR Target	Demonstrative	0.49	0.60	0.29	No

English	eng_ms_cl_pastreg_023	Cloze	Past Tense (Regular)	0.23	0.68	0.25	No
English	eng_ms_cl_poss_121	Cloze	Possessive 's	0.40	0.62	0.24	No
English	eng_ms_sr_pres3s_170	SR Target	Third Person Singular	0.38	0.71	0.24	No
English	eng_ms_cl_pastirreg_126	Cloze	Past Tense (Irregular)	0.50	0.62	0.22	No
English	eng_ms_cl_pres3s_122	Cloze	Third Person Singular	0.35	0.68	0.20	No
English	eng_ms_cl_pastreg_016	Cloze	Past Tense (Regular)	0.44	0.52	0.19	No
English	eng_ms_sr_pres3s_179	SR Target	Third Person Singular	0.30	0.74	0.17	No
English	eng_ms_cl_passive_049	Cloze	Passive	0.09	0.76	0.14	No
English	eng_ms_sr_wo_174	SR Verbatim	SR Verbatim	0.39	0.54	0.12	No
English	eng_ms_sr_wo_207	SR Verbatim	SR Verbatim	0.08	0.45	0.03	No
Spanish	spn_ms_sr_rel_069	SR Target	Relative Pronoun	0.78	0.71	0.75	Yes
Spanish	spn_ms_sr_artposs_282	SR Target	Possessive Article	0.65	0.73	0.66	Yes
Spanish	spn_ms_sr_rel_204	SR Target	Relative Pronoun	0.46	0.76	0.63	No
Spanish	spn_ms_sr_subj_206	SR Target	Subjunctive	0.55	0.71	0.61	Yes
Spanish	spn_ms_cl_adjfp_157	Cloze	Plural Adjective	0.68	0.70	0.58	Yes
Spanish	spn_ms_cl_subj_031	Cloze	Subjunctive	0.57	0.68	0.58	Yes
Spanish	spn_ms_sr_conj_236	SR Target	Conjunction	0.60	0.70	0.58	Yes
Spanish	spn_ms_sr_neg_242	SR Target	Negative	0.74	0.62	0.58	Yes
Spanish	spn_ms_sr_pret_278	SR Target	Preterite	0.65	0.68	0.53	Yes
Spanish	spn_ms_sr_cond_176	SR Target	Conditional	0.51	0.69	0.49	Yes
Spanish	spn_ms_sr_prep_203	SR Target	Preposition	0.59	0.64	0.49	Yes
Spanish	spn_ms_sr_pret_070	SR Target	Preterite	0.76	0.48	0.47	Yes
Spanish	spn_ms_cl_relclause_129	Cloze	Relative Clause	0.85	0.55	0.46	Yes
Spanish	spn_ms_sr_pret_093	SR Target	Preterite	0.78	0.53	0.46	Yes
Spanish	spn_ms_sr_rel_071	SR Target	Relative Pronoun	0.64	0.52	0.46	Yes
Spanish	spn_ms_sr_subj_173	SR Target	Subjunctive	0.57	0.66	0.46	Yes
Spanish	spn_ms_sr_neg_279	SR Target	Negative	0.82	0.57	0.44	Yes
Spanish	spn_ms_sr_prep_073	SR Target	Preposition	0.63	0.50	0.41	Yes
Spanish	spn_ms_sr_prep_281	SR Target	Preposition	0.76	0.59	0.41	No
Spanish	spn_ms_sr_prep_067	SR Target	Preposition	0.22	0.71	0.39	No
Spanish	spn_ms_sr_adv_200	SR Target	Adverb	0.69	0.55	0.36	No

Spanish	spn_ms_cl_imperf_151	Cloze	Imperfect	0.28	0.68	0.34	No
Spanish	spn_ms_cl_subj_028	Cloze	Subjunctive	0.62	0.53	0.31	No
Spanish	spn_ms_sr_prep_089	SR Target	Preposition	0.42	0.47	0.29	No
Spanish	spn_ms_sr_wo_244	SR Verbatim	SR Verbatim	0.55	0.44	0.29	No
Spanish	spn_ms_sr_wo_179	SR Verbatim	SR Verbatim	0.53	0.55	0.24	No
Spanish	spn_ms_sr_ioc_201	SR Target	Indirect Object Clitic	0.11	0.60	0.19	No

References

Albudoor, N., & Peña, E.D. (2021). Factors Influencing USA Speech and Language Therapists' Use of Technology for Clinical Practice. *International Journal of Language & Communication Disorders, 56*(3), 567-582.

Albudoor, N., Peña, E.D., & Bedore, L.M. (2019, November). *Diagnosing language impairment in bilingual children: speech recognition vs. human scoring* [Presentation]. University of California, Irvine Digital Learning Lab, Irvine, CA.

Allen, S. E. M., Genesee, F. H., Fish, S. A., & Crago, M. B. (2002). Patterns of code mixing in English-Inuktitut bilinguals. In *Proceedings of the 37th annual meeting of the Chicago Linguistic Society* (Vol. 2, pp. 171-188). Chicago, IL: Chicago Linguistic Society.

American Speech-Language-Hearing Association. (2021). *Bilingual Service Delivery* (Practice Portal). www.asha.org/Practice-Portal/Professional-Issues/Bilingual-Service-Delivery/

American Speech-Language-Hearing Association. (2018). *Demographic profile of ASHA members providing bilingual services, year-end 2018*. American Speech-Language-Hearing Association. http://www.asha.org.

Anaya, J. B., Peña, E. D., & Bedore, L. M. (2016). Where Spanish and English come together: A two dimensional bilingual approach to clinical decision making. *Perspectives of the ASHA special interest groups, 1*(14), 3-16.

Anderson, S. A., Hawes, D. J., & Snow, P. C. (2016). Language impairments among youth offenders: A systematic review. *Children and Youth Services Review, 65*, 195-203.

Arias, G., & Friberg, J. (2017). Bilingual language assessment: Contemporary versus recommended practice in American schools. *Language, Speech, and Hearing Services in Schools, 48*(1), 1-15.

Audacity Team (2020): Audacity (Version 2.3.2) [Computer program].

Bedore, L. M., Peña, E. D., Anaya, J. B., Nieto, R., Lugo-Neris, M. J., & Baron, A. (2018). Understanding disorder within variation: Production of English grammatical forms by English language learners. *Language, Speech, and Hearing Services in Schools*, *49*(2), 277-291.

Bedore, L. M., Peña, E. D., Gillam, R. B., & Ho, T. H. (2010). Language sample measures and language ability in Spanish-English bilingual kindergarteners. *Journal of Communication Disorders*, *43*(6), 498–510.

Bedore, L. M., Peña, E. D., Summers, C. L., Boerger, K. M., Resendiz, M. D., Greene, K., Bohman, T. M., Gillam, R. B. (2012). The measure matters: Language dominance profiles across measures in Spanish-English bilingual children. *Bilingualism*, *15*(3), 616–629.

Blom, E. (2010). Effects of input on the early grammatical development of bilingual children. *International Journal of Bilingualism*, *14*(4), 422–446.

Bolt, S. E., & Thurlow, M. L. (2004). Five of the most frequently allowed testing accommodations in state policy: Synthesis of research. *Remedial and special education, 25*(3), 141-152.

Boster, J. B. and McCarthy, J. W., 2018, Lost in translation: Understanding students' use of social networking and online resources to support early clinical practices. A national survey of graduate speech-language pathology students. *Education and Information Technologies*, *23*(1), 321–340.

Bracken, B. A., & McCallum, R. S. (1998). *Universal Nonverbal Intelligence Test*. Riverside Publishing Company.

Brinson, W., Cook, H., & Wellons, R. (2020). *A Systematic Review of Diagnostic Test Accuracy for Identifying Developmental Language Disorder in Bilingual Children* [Poster Presentation]. University of North Carolina at Chapel Hill, Department of Allied Health Sciences Division of Speech and Hearing Sciences Student Research Day, Chapel Hill, NC. https://doi.org/10.17615/srkb-pc16

Brownell, R. (2000). *Expressive One-Word Picture Vocabulary Test–Third Edition*. Academic Therapy Publications.

Brownell, R. (2001). *Expressive One-Word Picture Vocabulary Test: Spanish-Bilingual Edition*. Academic Therapy Publications.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Collins, B. A., O'Connor, E. E., Suárez-Orozco, C., Nieto-Castañon, A., & Toppelberg, C. O. (2014). Dual language profiles of Latino children of immigrants: Stability and change over the early school years. *Applied Psycholinguistics*, *35*(3), 581–620.

De Houwer, A. (2007). Parental language input patterns and children's bilingual use. *Applied Psycholinguistics, 28*(3), 411-424.

De Houwer, A. (2017). Bilingual language input environments, intake, maturity and practice. *Bilingualism: Language and Cognition, 20*(1), 19-20.

De Houwer, A. (2018). The role of language input environments for language outcomes and language acquisition in young bilingual children. In D. Miller, F. Bayram, J. Rothman, & L. Serratrice (Eds.), *Bilingual cognition and language: The state of the science across its subfields* (pp. 127-153). John Benjamins Publishing Company.

de Villiers, J., Iglesias, A., Golinkoff, R., Hirsh-Pasek, K., Wilson, M. S., & Nandakumar, R. (2021) Assessing dual language learners of Spanish and English: Development of the QUILS: ES. *Revista de Logopedia, Foniatría y Audiología*.

Dollaghan, C. A. (2007). *The Handbook for Evidence-Based Practice in Communication Disorders*. Paul H Brookes Publishing.

Dragon Systems, Inc. (1997). Dragon: NaturallySpeaking, [computer program]. SeanSoft.

Ebel, R.L., & Frisbie, D.A. (1986). *Essentials of educational measurement*. Prentice-Hall.

Ebert, K. D., & Kohnert, K. (2016). Language learning impairment in sequential bilingual children. *Language Teaching, 49*(3), 301.

Flores, G., Laws, M., Mayo, S., Zuckerman, B., Abreu, M., Medina, L., & Hardt, E. (2003). Errors in medical interpretation and their potential clinical consequences in pediatric encounters. *Pediatrics, 111*, 6-14.

Hoff, E., Core, C., Place, S., Rumiche, R., Señor, M., & Parra, M. (2012). Dual language exposure and early bilingual development. *Journal of Child Language*, *39*(1), 1–27.

Gillam, R. B., & Pearson, N. A. (2004). TNL: *Test of Narrative Language*. Pro-ed.

Gillam, R. B., Peña, E. D., Bedore, L. M., Bohman, T. M., & Mendez-Perez, A. (2013). Identification of specific language impairment in bilingual children: I. Assessment in English. *Journal of Speech, Language, and Hearing Research, 56*(6), 1813-1823.

Gillam, R. B., Peña, E.D., Bedore, L.M., & Pearson, N. A. (2010). *Test of Narrative Language, Spanish Experimental Version* (unpublished test).

Glogowska, M., Roulstone, S., Enderby, P., & Peters, T. J. (2000). Randomized controlled trial of community based speech and language therapy in preschool children. *British Medical Journal, 321*(7266), 923–928.

Goldman, R. (2015). *Goldman-Fristoe Test of Articulation, Third Edition*. AGS.

Goldschneider, J. M., & DeKeyser, R. M. (2001). Explaining the "natural order of L2 morpheme acquisition" in English: A meta-analysis of multiple determinants. *Language learning*, *51*(1), 1-50.

Google. (2020). *Google Cloud Speech-to-Text (Version 1) Documentation*. Google Cloud. https://cloud.google.com/speech-to-text/docs/

Gutiérrez-Clellen, V., Restrepo, A., & Simon-Cereijido, G. (2006). Evaluating the discriminant accuracy of a grammatical measure with Spanish-speaking children. *Journal of Speech, Language, and Hearing Research, 49*, 1209–1223.

Gutierrez-Clellen, V. F., Simon-Cereijido, G., & C. Wagner (2008). Bilingual children with language impairment: A comparison with monolinguals and second language learners. *Applied Psycholinguistics, 29*(1), 3−19.

Hair, A., Ballard, K. J., Ahmed, B., & Gutierrez-Osuna, R. (2019, October). Evaluating Automatic Speech Recognition for Child Speech Therapy Applications. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 578-580). ACM.

Lugo-Neris, M. J., Peña, E. D., Bedore, L. M., & Gillam, R. B. (2015). Utility of a language screening measure for predicting risk for language impairment in bilinguals. *American Journal of Speech-language Pathology / American Speech-Language-Hearing Association*, *24*(3), 426–437.

MacArthur, C. A., & Cavalier, A. R. (2004). Dictation and speech recognition technology as test accommodations. *Exceptional Children, 71*(1), 43-58.

Meisel, J. M. (2007). The weaker language in early child bilingualism: Acquiring a first language as a second language? *Applied Psycholinguistics*, *28*(3), 495–514.

Morgan, P. L., Farkas, G., Hillemeier, M. M., Mattison, R., Maczuga, S., Li, H., & Cook, M. (2015). Minorities are disproportionately underrepresented in special education: Longitudinal evidence across five disability conditions. *Educational Researcher, 44*(5), 278-292.

Norbury, C. F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., Vamvakas, G., & Pickles, A. (2016). The impact of nonverbal ability on prevalence and clinical presentation of language disorder: evidence from a population study. *Journal of Child Psychology and Psychiatry, 57*(11), 1247-1257.

Oppenheim, G. M., Griffin, Z., Peña, E. D., & Bedore, L. M. (2020). Longitudinal evidence for simultaneous bilingual language development with shifting language dominance, and how to explain it. *Language Learning, 70*(S2), 20-44.

Orgassa, A. & Weerman, F. (2008). Dutch gender in specific language impairment and second language acquisition. *Second Language Research, 24*(3), 333−364.

Paradis, J. & Crago, M. (2000). Tense and temporality: A comparison between children learning a second language and children with SLI. *Journal of Speech, Language, and Hearing Research, 43*(4), 834−847.

Peña, E. D., & Bedore, L. M. (2011). It takes two: Improving assessment accuracy in bilingual children. *ASHA Leader, 16*(13), 20.

Peña, E. D., Bedore, L. M., & Griffin, Z. (2010). Cross-language outcomes of typical and atypical development in bilinguals: National Institute of Deafness and Other Communication Disorders.

Peña, E. D., Bedore, L. M., Iglesias, A., Gutierrez-Clellen, V. F., & Goldstein, B. A. (in development). *Bilingual English Spanish Oral Screener (BESOS), Experimental Version* (unpublished test).

Peña, E. D., Bedore, L. M., Iglesias, A., Gutierrez-Clellen, V. F., & Goldstein, B. A. (in development). *Bilingual English Spanish Assessment-Middle Elementary (BESA-ME), Experimental Version* (unpublished test).

Peña, E. D., Bedore, L. M., & Kester, E. S. (2016). Assessment of language impairment in bilingual children using semantic tasks: Two languages classify better than one. *International Journal of Language & Communication Disorders, 51*(2), 192–202.

Peña, E. D., Bedore, L. M., Lugo-Neris, M. J., & Albudoor, N. (2020). Identifying developmental language disorder in school age bilinguals: semantics, grammar, and narratives. *Language Assessment Quarterly*, 1-18.

Peña, E. D., Gutiérrez-Clellen, V. F., Iglesias, A., Goldstein, B., & Bedore, L. M. (2010). *Bilingual Input Output Survey (BIOS) and Inventory to Assess Language Knowledge (ITALK), Experimental Version* (unpublished test).

Peña, E. D., Bedore, L. M., Shivabasappa, P., & Niu, L. (2018). Effects of divided input on bilingual children with language impairment. *International Journal of Bilingualism*, 136700691876836. https://doi.org/10.1177/1367006918768367

Peña, E. D., Gutiérrez-Clellen, V. F., Iglesias, A., Goldstein, B., & Bedore, L. M. (2018). *Bilingual English Spanish Assessment*. Brookes Publishing Co.

Pettit, B., & Western, B. (2004). Mass imprisonment and the life course: Race and class inequality in US incarceration. *American Sociological Review, 69*(2), 151-169.

Plante, E., & Vance, R. (1994). Selection of preschool language tests: A data-based approach. *Language, Speech, and Hearing Services in Schools, 25*(1), 15–24.

Pratt, A. S., Adams, A. M., Peña, E. D., & Bedore, L. M. (2020). Exploring the use of parent and teacher questionnaires to screen for language disorders in bilingual children. *Topics in Early Childhood Special Education*.

Pratt, A. S., Anaya, J. B., Ramos, M., Peña, E. D., & Bedore, L. M. (in review). From a distance: Comparison of in-person and virtual assessment with adult-child dyads from linguistically diverse backgrounds. *Language, Speech, and Hearing Services in Schools*.

Roberts, M. Y., & Kaiser, A. P. (2015). Early intervention for toddlers with language delays: A randomized controlled trial. *Pediatrics, 135*, 686–693.

Royal College of Speech and Language Therapists. (2021) *Bilingualism Overview.* *https://www.rcslt.org/speech-and-language-therapy/clinical-information/bilingualism*

Sabu, K., & Rao, P. (2018). Automatic assessment of children's oral reading using speech recognition and prosody modeling. *CSI Transactions on ICT, 6*(2), 221-225.

Saenz, T. I., & Langdon, H. W. (2019). Speech-language pathologists' collaboration with interpreters: Results of a current survey in California. *Translation & Interpreting, 11*(1), 43-62.

Santhanam, S. P., Gilbert, C. L., & Parveen, S. (2019). Speech-language pathologists' use of language interpreters with linguistically diverse clients: A nationwide survey study. *Communication Disorders Quarterly, 40*(3), 131-141.

Spille, C., Kollmeier, B., & Meyer, B. T. (2018). Comparing human and automatic speech recognition in simple and complex acoustic scenes. *Computer Speech & Language, 52*, 123-140.

Sullivan, A. L., & Bal, A. (2013). Disproportionality in special education: Effects of individual and school variables on disability risk. *Exceptional Children, 79*(4), 475-494.

Thompson, S., Blount, A., & Thurlow, M. (2002). A Summary of Research on the Effects of Test Accommodations: 1999 through 2001. Technical Report.

Thurlow, M. L., & Johnson, D. R. (2011). The high school dropout dilemma and special education students. University of California, Santa Barbara, Santa Barbara, CA.

U.S. Census Bureau (2019). *American Community Survey*. U.S. Census Bureau. https://www.census.gov/programs-surveys/acs

Wood, D.A. (1960). *Test construction: Development and interpretation of achievement tests*. Charles E. Merrill Books, Inc.

Wood, L., Kiperman, S., Esch, R. C., Leroux, A. J., & Truscott, S. D. (2017). Predicting dropout using student-and school-level factors: An ecological perspective. *School Psychology Quarterly, 32*(1), 35.

Xu, Y., Wang, D., Collins, P., Lee, H., & Warschauer, M. (2021). Same benefits, different communication patterns: Comparing Children's reading with a conversational agent vs. a human partner. *Computers & Education, 161*, 104059.